日本データベース学会和文論文誌
Vol. 22-J, Article No. 2, 2024 年 3 月

部分構造を考慮した化合物分散表現の食材分類タスクにおける効果

吉丸 直希 [1]　木村 優介 [2]　楠 和馬 [3]
波多野 賢治 [4]

近年，機械学習タスクの実行に有用な特徴表現を構築するための表現学習の知見をもとに，化学の分野に活かすケモインフォマティクスに注目が集まっている．その中に，化合物をベクトル表現する取組みがあり，食材が複数の化合物から構成されている事実から，食メディア分野での応用が期待されている．既存手法では，化合物の原子団である部分構造の有無だけで化合物のベクトルを表現している．しかし，この単純な表現方法では，部分構造間における関係の複雑さが表現されておらず，化合物の表現としては十分ではない可能性がある．そこで本論文では，食材に関連する化合物データを用いて，部分構造の関係性を考慮した化合物のベクトル表現法を提案する．提案した化合物表現を用いて食材を表現したところ，食材の分類タスクにおいて既存手法より精度が向上したことを確認できた．

1　はじめに

食は人間にとって最も基本的な活動の一つであり，食にまつわる知識は生活する上では無くてはならない重要なものである．そのため食に関する研究は，人間の基本的な生活を支援するために昔から行われてきており，栄養学の分野に限らず，医学や情報学など幅広い分野で扱われている．特に，情報学においては，これまで書籍やアンケート等の古典的な方法でレシピデータの収集を行い，それを活用する研究が行われてきたが，近年ではレシピ提供サイトの充実化が図られ，さまざまなレシピが日々生成・蓄積され，これまで以上に容易にレシピデータを用いた研究が行える環境が整ってきている．

レシピデータには非常に多くの種類のメディアが含まれている．分かりやすいところでいえば，文字列により表現される食材や使用器具といったテキストデータに始まり，レシピで使用される食材の使用量や含有栄養素量などの数値データ，調理手順を示すための画像や映像，音声データなど多岐にわたる．こうしたデータを活用すれば，食文化の把握や調理技術の習得・伝達・継承，そして，食事に基づく健康管理などの分野において応用研究が可能となっており，こうした研究は食メディア研究（Food Computing）と呼ばれている [1].

[1] 学生会員　同志社大学大学院文化情報学研究科
yoshimaru@mil.doshisha.ac.jp
[2] 学生会員　同志社大学大学院文化情報学研究科
kimura@mil.doshisha.ac.jp
[3] 学生会員　同志社大学大学院文化情報学研究科
kusu@mil.doshisha.ac.jp
[4] 正会員　同志社大学文化情報学部
khatano@mail.doshisha.ac.jp

食メディア研究における課題の一つに，食材表現の構築に関する研究がある．これは食材を計算機で扱える形式で表現することを指しているが，例えばリンゴという食材を「りんご」という文字列で表現するということではない．具体的には地域文化の違いの解明を目的に，文化的な生活様式に対するアンケート調査データから食材表現を作成する研究 [2] や，食材画像における局所特徴を認識可能にする食材表現を構築する研究 [3]，健康維持を目的に食事履歴から食材に対する個人の嗜好性をベクトル表現する研究 [4] などのことである．それぞれの表現形式は，それぞれ個別のタスクを解くために用いられる食材の表現形式だが，情報学的には，食材は本来，正確にさまざまな分野で汎用的に使用できるような表現形式をとるべきである．

このような傾向は化学の研究分野においても見られ，情報学分野で行われていた文書中に出現する文字列同士の共起関係を，化合物中に含まれている元素同士の共起関係や元素同士の結合の種類などを利用して，化合物をベクトル表現する研究が行われている [5]．この表現方式を利用することで，化学分野では同じ特徴をもつ化合物をトピックモデルを用いて捉える研究 [6] が行われており，この表現を基礎に新たな化合物の発見や新薬の生成などに活用し，このような分野全体を化学情報学（ケモインフォマティクス）と呼び盛んに研究が行われている．しかし前述のように，それぞれの個別のタスクを解くためにさまざまな化合物表現法が存在しているに過ぎず，化合物の特徴を捉えた汎用的に使用できる表現形式を構築する要望は大きい．

そこで本論文では，食材が複数の化合物から形成されている点に注目し，情報学でこれまで提案されてきた食材表現法ではなく，化学情報学で用いられている食材表現法を適用することで，従来よりも正確に食材の表現が可能かどうかを確認する．この方法を活用することで，レシピ内に出現する語の共起関係から構築されるという食材表現法より，化学的見地に立った食材表現を実現することができるため，これ以後，食材に関わるさまざまなタスクでこの表現を使用する事ができる可能性がある．

2　関連研究

本節では，本研究に関連する食材や化合物のデータを使って，それぞれの食材や化合物のベクトルデータを生成する研究事例を紹介する．情報学の分野ではその表現を分散表現と呼ぶが，これら分散表現は多くのデータから機械学習の手法を用いて構築されるため，このような方法を表現学習とも呼ぶ．

2.1　食材の分散表現

個々の食材をベクトルの形式で表現する研究として挙げられる事例として挙げるべき研究は FlavorGraph を構築する研究である [7]．FlavorGraph が開発された目的は，相性のよい食材ペアを発見することであるが，既存の研究で行われてきた食材同士の組合せパタンをルール化する研究では，あらゆる食材の組合せを考慮することができず，網羅的に相性のよい食材ペアを発見できないという問題が指摘されていた*1．FlavorGraph はこの

*1 Emerging Technology from the arXiv on MIT Technology Review, “Flavor Networks Reveal Universal Principle Behind Successful Recipes,” https://

問題に対処すべく，食材とそれに関連する化合物から図 1 のような異種グラフを構築し，この構造を使って各食材の固定長ベクトルを算出するというものである．固定長ベクトルの算出には Metapath2vec [8] と呼ばれるグラフ構造の一部を辿ったノードのパスを機械学習の入力とする方法であり，Node2vec [9] や LINE [10] などに類似した手法である．これらと異なる点は，ノードのパス生成ルールを異種グラフの状況ごとに定めることで，異種グラフの性質を固定長ベクトルに反映することができることである．FlavorGraph では，食材がいくつかの化合物から構成されていることを鑑み，特に食材と化合物の関係を綿密に表現できるようにパスを構築しており，その構築方法は図 2 で示す通りである．これはグラフ上のノードを無作為にウォークするランダムウォークにより生成されるものをパスと呼ぶことであり，FlavorGraph の場合は食材と化合物の間を一定のルールに従ってウォークする．

また，化合物の表現を化合物全体を一つの表現で表す分子フィンガープリント（Molecular Fingerprints）記法を用いて表現 [11] し，Metapath2vec の中間層に組み込むことで，食材だけでなく化合物自体も類似した化合物同士が似たものとして扱われる[*2] ような工夫も行っている．これらの工夫により，レシピデータ内のデータだけを用いた既存研究では見出せなかった，相性のよい食材関係を正確な化合物表現を介して予測できるようになっている．分子フィンガープリントの方式は，現在もさまざまな方法が提案されている [12]．FlavorGraph は情報学分野で考えられた異種グラフ構造に基づく固定長ベクトルの算出手法を化学分野と関連させたものであるが，使用する分子フィンガープリントが正しく化合物を表現できていなければ，食材分散表現の構築は正しく行えない可能性がある．食材は複数の化合物から構成されており，食材内の化合物はランダムではない一定の規則で複雑に絡み合っている [13] ことが言われている．このことから，FlavorGraph で使用されている分子フィンガープリントは，特定の部分構造の有無であり，複雑な性質を表現できていない可能性があり，食材分散表現の構築する手段として改善の余地があると言える．

2.2 化合物の分散表現

化学情報学において，最も基本的なことは分子を計算機で扱える形式で表現することであり，2.1 節でも述べたように多くの分子フィンガープリントの方法が提案されている．これは，世の中には未だ発見されていない化合物も多く，その構造が明らかになっていないことから行われており，多様な性質を持つ化合物を計算機で表現することを目標としている．一度，化合物を計算機で表現することができれば，それらの類似性検索や特定構造の検索，分子構造の予測などのタスクが実行可能となるため，化学情報学ではさまざまな目的に応じて分子フィンガープリントで化合物を表現し，そのデータを用いて機械学習することで，多くの化学分野における課題解決に取り組んでいる．

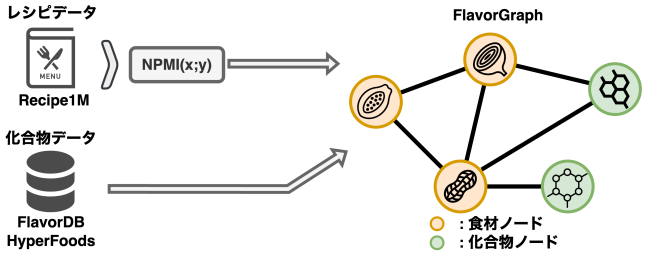

図 1: FlavorGraph の構築手順

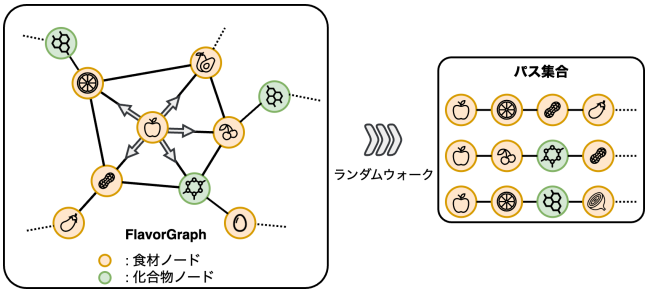

図 2: パス生成のイメージ図

情報学の分野の成功を化学の世界に持ち込んだ研究として挙げるべきは Mol2vec である [5]．これは文書に含まれている単語を計算機で扱う際にその出現頻度という数値で計算されていたものがベクトルで表現できるような技術 word2vec [14] が提案され，その成功にインスパイアされた方法である．その仕組みは，化合物を word2vec における文，部分構造を文を構成する語と捉え，化合物の部分構造のベクトル表現を得るというものである．word2vec をベースに文書のベクトル化を行う doc2vec [15] が提案されているように，化合物全体のベクトル表現も容易に得ることができる．

本論文では，2.1 節で述べた FlavorGraph における化合物表現に分子フィンガープリントを直接使用するのではなく，その分子フィンガープリントを入力として得られた各化合物の Mol2vec による出力を用いて，食材分類タスクの評価を行う．これにより，化学情報学で行われている化合物表現手法が，食材を表現する際に効果示すかどうかを確認することができる．

3 提案手法

2 節で述べたように，FlavorGraph を用いた食材表現を用いた場合，単純に分子フィンガープリントを利用した結果をそのまま用いているに過ぎないため，食材間の関係性を化合物を介して表現することができず，その結果，出力されている食材表現のベクトルが正確に生成されていない可能性がある．そこで本論文では，化学情報学分野で行われた Mol2vec を用いて食材分散表現を構築するために既存手法（FlavorGraph）を改良する方法を提案する．具体的には，FlavorGraph から Metapath2vec を用いて食材分散表現を構築する際，既存手法では 0, 1 で表現されたの

www.technologyreview.com/2017/04/20/152437/flavor-networks-reveal-universal-principle-behind-successful-recipes/, 2023 年 12 月 14 日閲覧.

[*2] 化合物の類似性は，研究の目的や用いるデータセットによって異なるため，使用する分子フィンガープリントの方法は異なるが，FlavorGraph では原子などの要素数や環構造の種類，部分構造の結合条件などの有無を 0 もしくは 1 で表現した 881 次元のバイナリベクトル（PubChem Fingerprints）で表現している．

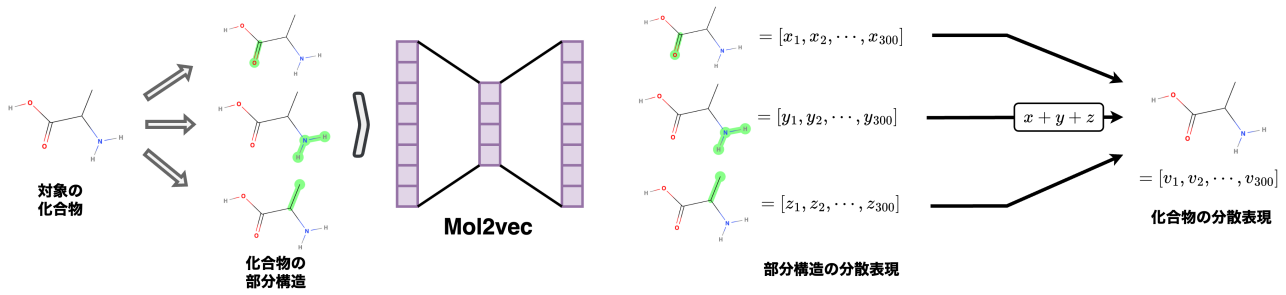

図 3: 化合物表現の構築

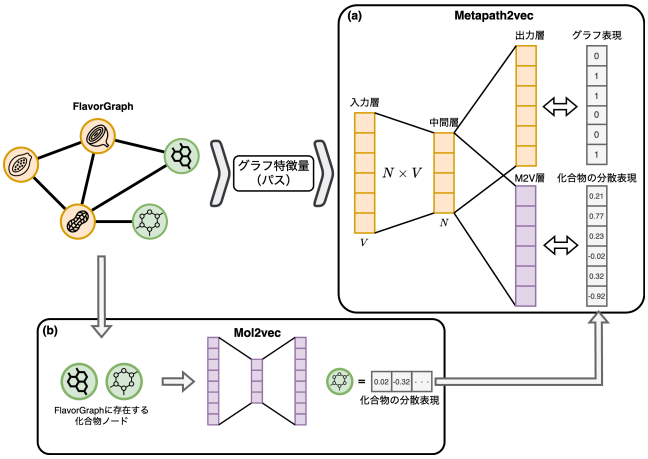

図 4: 食材表現の構築手順

Pubchem fingerprint を取り入れていたが，本手法では Mol2vec で学習した連続値の化合物表現を利用する．これらによって作成された食材分散表現を，既存手法である FlavorGraph と比較し，より精緻に食材ベクトルを構築できたどうかを検証する．

食材表現ベクトルの生成手順の概要は以下の通りである．

1. 食材を構成するとされている化合物を FlavorDB[*3] より抽出
2. 抽出した化合物の名称から，世界最大級の化学分子データベースである PubChem [*4] から，化合物を一意に同定
3. 化合物に付与されている PubChem ID（CID）をキーとして分子フィンガープリントを取得
4. 得られた分子フィンガープリントを使用して Mol2vec で学習し，化合物のベクトル表現を取得
5. FlavorGraph の学習時に使用される Metapath2vec の中間層に，直上のステップで取得した化合物表現ベクトルを活用し，食材の関係性と化合物の表現を同時に予測しながら学習するマルチタスク学習を実行
6. マルチタスク学習のメイン学習タスクの結果として，求める食材表現ベクトルを取得

[*3] Ganesh Bagler, “FlavorDB", https://cosylab.iiitd.edu.in/flavordb/, 2023 年 12 月 14 日閲覧.

[*4] National Center for Biotechnology Information, “PubChem,” https://pubchem.ncbi.nlm.nih.gov/, 2023 年 12 月 14 日閲覧.

ステップ（3）の分子フィンガープリントの取得には，Python ライブラリである RDKit[*5] を用い，使用したフィンガープリント生成アルゴリズムは Morgan Fingerprints（MF）[16] である．MF は各原子からそれ以外の原子がどれほど離れているかを数え，その値を結合距離を設定し，その距離ごとに部分構造を分けて扱うことで分子全体を表現する方法であり，近年の化学情報学においては頻繁に使用されている方法である [17]．本論文においてもその傾向に倣い， MF で分子を表現する方法を採用した上で結合距離を 1 と設定し，図 3 内の緑部分で表される部分構造を表す分子フィンガープリントを Mol2vec の入力として使用している．

ステップ（4）の Mol2vec を用いた学習では，ステップ（4）で得られた部分構造を word2vec における「単語」，化合物全体を「文」と見做すことで学習が行われ，化合物の共起関係を学習することで，類似した部分構造同士がベクトル空間内で近傍に付置されるような出力結果が得られる．各部分構造のベクトル表現が得られた後，再度，化合物単位で集計することで，その化合物のベクトル表現も得ることができる．

ステップ（5）のマルチタスク学習の詳細は図 4 であり，FlavorGraph における食材と化合物の関係性を予測するメイン学習タスク (図 4（a）) と，Mol2vec により構築された化合物分散表現を予測するサブ学習タスク (図 4（b）) の二つから構成されている．食材と化合物の関係性を予測するタスクでは FlavorGraph で得られたパスを入力し，実際の FlavorGraph でのノード関係との差異を小さくするよう学習を進める．なお，食材分散表現として得るベクトルの次元は FlavorGraph と同様に 300 次元としており，図 4 の Metapath2vec の学習における N の部分が食材分散表現として得られる．

以上の処理を行うことで，食材と化合物との関係性がレシピ内に現れていなくても，得られる食材表現が化合物の表現により補完され，従来より正確に表現できると考えられる．

4 評価実験

提案手法の有効性を評価するために，先行研究（FlavorGraph）により作成される食材分散表現との比較実験を行った．先行研究の手法ではマルチタスク学習のサブ学習タスクで PubChem Fingerprints を使用していたが，提案手法では Mol2vec により生

[*5] RDKit, http://www.rdkit.org/, 2023 年 12 月 14 日閲覧.

表 1: 分類結果

手法	SVM		MLP	
	macro F_1	micro F_1	macro F_1	micro F_1
既存手法 (PubChem Fingerprints)	0.234	0.311	0.234	0.311
提案手法 (Mol2vec)	**0.302**	**0.377**	**0.299**	**0.361**

成される化合物表現を用いている．

4.1 異種グラフの構築

比較実験を行うためには FlavorGraph が行ったスクレイピングを実行し同じデータを使用する必要があったが，食材と薬化合物との関連を収集した Hyperfoods [18] のデータで Mol2vec に適応するために必要な化合物の構造に関する情報を取得することが困難であった．そのため，FlavorGraph で使用しているレシピデータセットの Recipe1M [19] と，食材と化合物の関係情報を持つデータベースである FlavorDB [20] を用いて FlavorGraph と同程度の規模の異種グラフを構築した．食材については 2 節での FlavorGraph と同様に大規模なレシピデータセットである Recipe1M [19] と，食材と化合物の関係情報を持つデータベースである FlavorDB [20] から構築した．その際，以下の二つの条件をもとに食材同士にエッジを張るかを判断する．

- 各食材が Recipe1M に 20 回以上登場し，かつ両者が同じレシピに 5 回以上登場する
- NPMI スコアが 0.25 以上

ここで NPMI スコアは以下の式 (1) で表される共起を表す値であり，-1 から 1 の間の値を取り，値が大きいほどレシピに同時に登場しやすいことを表す．

$$\mathrm{NPMI}(x;y) = \left(\log \frac{p(x,y)}{p(x)p(y)}\right) \Big/ -\log p(x,y) \tag{1}$$

これらの条件により，全レシピの中で 1, 2 回共に登場するといった極端に少ない食材ペアを除外することができる．それぞれのデータセットから得られる食材を照合しグラフを構築する．以上の手法で構築された異種グラフのノード，エッジの種類と数は，表 3 に示す通りである．

4.2 評価方法

既存手法，提案手法それぞれで生成された食材のベクトル表現が，食材表現として相応しいものなのかを評価するために，生成した食材表現ベクトルを用いたカテゴリ分類を行った．ここで扱うカテゴリは食材に対して与えられるものであり，肉類や魚介類，穀物などのことである．それぞれの手法で得られた食材のベクトル表現によって分類された食材が，実際の食材分類の結果を再現できていればいるほど，分散表現が精緻化されたと判断できる．これはカテゴリのデータがない状態で学習された食材分散表現が，食材のカテゴリを再現できたことを表せており，より食材の本質に近付いたと言えるからである．

評価用データのカテゴリとその内訳（食材の数）は，表 2 に示す通りである．これらのデータは FlavorGraph の研究の際に使用された人手で作成された評価データであり，食材と化合物の関係をまとめたデータベースである FlavorDB [20] から作成されている．

表 2: 評価用データ

カテゴリ	データ数
ベーカリー・デザート・スナック	38
飲料	18
アルコール飲料	26
穀物・豆類	56
乳製品	47
精油・油脂	14
小麦粉	4
果実	57
菌類	6
肉・動物製品	30
ナッツ・種子	27
植物・野菜	147
ソース・パウダー・ドレッシング	67
魚介類	35
スパイス	24
料理	16
その他	4

食材ベクトルの分類を行うために多クラス分類手法用いることが考えられるが，一つの分類手法だけで評価した場合，扱うデータセットに過剰に適合している可能性を排除できず，食材分散表現の性能を正確に測ることができない．これらの課題を解消するために，多クラス分類手法として有名な多クラスサポートベクトルマシン（SVM）と多層パーセプトロン（MLP）を用いた．これらの分類手法は異なるアルゴリズムで分類を行うため，片方の分類手法だけで精度が高くなるといった偏りを解消することが可能になる．評価指標は表 2 に示されるように使用したデータセットが不均衡であることを考慮し，macro F_1, micro F_1 で評価する．これらは 2 値分類における，適合率 P と再現率 R の調和平均を取った F_1 スコアを多クラス分類に拡張した指標であり，以下の式 (2)，(3) で定義される．

$$\text{macro } F_1 = \frac{1}{N}\sum_{k}^{N} \frac{2P_kR_k}{P_k + R_k} \tag{2}$$

$$\text{micro } F_1 = \frac{2\bar{P}\bar{R}}{\bar{P} + \bar{R}} \tag{3}$$

ここで P_k，R_k は各クラス k における適合率・再現率であり，$\bar{P}$,

(a) PubChem Fingerprints　　(b) Mol2vec (Morgan Fingerprints)

図 5: 化合物表現の分布比較

表 3: グラフ構成要素

構成要素	種類	数
ノード	食材	6,653
	化合物	1,559
エッジ	食材 - 食材	111,355
	食材 - 化合物	35,143

$\bar{R}$ は全クラスで平均を取った適合率・再現率である．micro F_1 は各カテゴリラベルでの F_1 スコアを平均したものであり，全体の分類性能を測ることができる．しかし，あるカテゴリでのデータが少なく，著しく精度が低い場合においても F_1 スコアが平均されるため，その状況が精度に現れない．そのため不均衡データにおいては macro F_1 も同時に用いられている．こちらは各クラスラベルでの P_k と R_k を平均した上で F_1 スコアを計算するため，各ラベルでの P_k と R_k がスコアに反映され，不均衡データに対応した指標として適している．

4.3 実験結果

多クラス分類手法に基づく食材の分類結果を表 1 に示す．それぞれの分類手法，評価指標において，提案手法に基づく食材表現ベクトルを用いた方が分類精度向上が見られた．したがって，Mol2vec を用いて化合物表現を学習させた結果を利用した食材表現ベクトルを使用した方が，食材分類タスクにおいて効果的であることが確認することができた．

4.4 考察

今回の提案手法で用いた Mol2vec による化合物表現は計 300 次元の表現であり，既存の PubChem Fingerprints による方法の 881 次元と比べると低次元で密なベクトルである．そこで，次元を少なくしたことによる分類精度への影響を確かめるため，次元圧縮した化合物表現を用いて再度分類を行った．具体的には 2.1 節の FlavorGraph で用いていた 881 次元で表される化合物表現を，入力層 881 次元，隠れ層 300 次元，出力層 881 次元の単純な Auto Encoder [21] を用いて次元圧縮した．そして隠れ層で計算される 300 次元の化合物表現を用いて再度実験を行う．これは提案手法が 300 次元で計算していることに起因しており，提案手法の精度向上に次元の少なさがどれほど影響を及ぼしたかを確認することができる．

表 4: 次元圧縮による精度比較 (MLP)

	881 次元	300 次元
micro F_1	0.311	**0.369**

分類結果が表 4 になる．他の分類手法においても表 4 の MLP での結果のように少なからずスコアが向上していた．そのため，化合物のベクトルが 881 次元から 300 次元によりなることにより精度が上がったため，化合物表現のベクトルが密になることによる食材表現への影響があることがわかる．

また化合物表現に関する比較として，それぞれの分布の違いを観察した．881 次元，300 次元それぞれの化合物表現を 2 次元に圧縮し，それらをプロットしたものが図 5 になる．結果，Mol2vec で得られた化合物表現の方がある一定のまとまりが見られた．部分構造の関係を学習した結果，化合物の細かな性質の違いを表現できたため，これらの違いが現れたと考えられる．

5 おわりに

本論文では，食材のベクトル表現を得るために，当該食材に関連する化合物の表現を活用する方法を提案した．その結果，食材分類タスクにおいて分類精度の向上が見られ，Mol2vec による化合物表現を活用することの有効性を確認することができた．

今後の課題としては，学習過程においての工夫が考えられる．今回は化合物表現を食材表現の学習に取り入れる二段階の仕組みであり，食材と化合物の関係を直接，食材表現に学習できていない．そのため，食材表現を学習する際と同時に Mol2vec などの化合物表現を学習する手法の検討が挙げられる．

謝辞

本研究の一部は，日本学術振興会科学研究費助成事業基盤研究 (A)19H01138，および基盤研究 (B)19H04218 の助成を受けて遂行された．ここに記して謝意を表す．

参考文献

[1] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, Vol. 52, No. 5, pp. 1–36, 2019.
[2] Xu Chen. Multilingual analysis of food words. Master's thesis, Universität Stuttgart, 2019.
[3] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[4] Ahmed A Metwally, Ariel K Leong, Aman Desai, Anvith Nagarjuna, Dalia Perelman, and Michael Snyder. Learning personal food preferences via food logs embedding. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2281–2286. IEEE, 2021.
[5] Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, Vol. 58, No. 1, pp. 27–35, 2018.
[6] Nadine Schneider, Nikolas Fechner, Gregory A Landrum, and Nikolaus Stiefl. Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *Journal of chemical information and modeling*, Vol. 57, No. 8, pp. 1816–1831, 2017.
[7] Donghyeon Park, Keonwoo Kim, Seoyoon Kim, Michael Spranger, and Jaewoo Kang. Flavorgraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. *Scientific reports*, Vol. 11, No. 1, pp. 1–13, 2021.
[8] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, p. 135–144, New York, NY, USA, 2017. Association for Computing Machinery.
[9] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 855–864, New York, NY, USA, 2016. Association for Computing Machinery.
[10] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, p. 1067–1077, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.
[11] D. Bajusz, A. Rácz, and K. Héberger. 3.14 - chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching. In Samuel Chackalamannil, David Rotella, and Simon E. Ward, editors, *Comprehensive Medicinal Chemistry III*, pp. 329–378. Elsevier, Oxford, 2017.
[12] Mahendra Awale, Ricardo Visini, Daniel Probst, Josep Arús-Pous, and Jean-Louis Reymond. Chemical space: big data challenge for molecular diversity. *CHIMIA International Journal for Chemistry*, Vol. 71, No. 10, pp. 661–666, 2017.
[13] David R Jacobs and Linda C Tapsell. Food synergy: the key to a healthy diet. *Proceedings of the Nutrition Society*, Vol. 72, No. 2, pp. 200–206, 2013.
[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.
[15] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR, 2014.
[16] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, Vol. 50, No. 5, pp. 742–754, 2010.
[17] Noel MO' Boyle, Roger A Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of cheminformatics*, Vol. 8, No. 1, pp. 1–14, 2016.
[18] Kirill Veselkov, Guadalupe Gonzalez, Shahad Aljifri, Dieter Galea, Reza Mirnezami, Jozef Youssef, Michael Bronstein, and Ivan Laponogov. Hyperfoods: Machine intelligent mapping of cancer-beating molecules in foods. *Scientific reports*, Vol. 9, No. 1, pp. 1–12, 2019.
[19] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3020–3028, 2017.
[20] Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi Nk, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, et al. Flavordb: a database of flavor molecules. *Nucleic acids research*, Vol. 46, No. D1, pp. D1210–D1216, 2018.
[21] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, Vol. 313, No. 5786, pp. 504–507, 2006.